Part 2: Description of Work

1. Summary

We are in a massive human-driven biodiversity extinction with uncertain consequences for Earth climate, life conditions and the stability of Earth (Figure 1). This rapid global change put us in an edge to take in science the risks to reduce the uncertainty related to the consequences of feedbacks between the Earth system and biodiversity (Figure 2). The rapid decline of biological diversity urges efforts from the scientific community to characterize and to identify key players: Biodiversity is sustained by processes acting across biological (i.e., from genes to traits and ecological networks). Yet, the fussion of heterogeneous data spanning different biological levels, taxa and ecosystems to decipher the interdependencies among biological levels and how such interdependences alter biodiversity response to rapid global change is still not in place. This project will develop a deep learning biodiversity platform based on processes to leverage data collected from several sources to map future scenarios of biodiversity decline in interdependent biological networks.

Statements of the goals

Our main goal is to pursue how interdependent biological levels affect biodiversity dynamics and its decline. Biodiversity research has been sistematically studied at one biological level. While splitting disciplines in many biological, temporal and spatial scales have produced an immense gain in detailed knowledge at each of the levels and scales studied, it might be insuffcent to understand the consequences of biodiversity dynamics when feedbacks between biological levels occur (Figures 1 and 2). We propose to study data-driven patterns within and between biological levels to decipher the role of interdependencies for the maintenance of biodiversity. Fussioning modern data analytics and theory in biodiversity research would help to contrast process-based scenarios of biodiversity decline with and without considering interdependences across biological levels (Figure 2).

Milestones

i) **Deep process-based learning platform for biodiversity research**: We will explore deep-learning networks to contrast patterns of biodiversity dynamics with and without feedbacks within and between the networks (Box 1 and Box 2).
ii) **Visualization tools**: We will communicate in public exhibitions the core patterns and processes governing interdependent networks across biological levels and its impact on biodiversity dynamics and current decline rates.

Significance of the project for data science

We will bring deep process-based learning networks to biodiversity research as a fundamental and applied tool to unfold the role of the interdependencies among biological levels for understanding biodiversity dynamics.

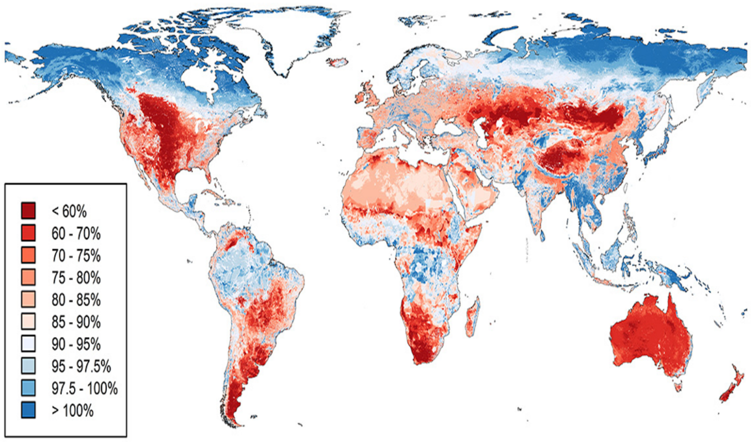


Figure 1: Biodiversity is declining globally at unprecedented rates. Map showing the remaining populations of native species across many taxa as a percentage of their original populations. Blue areas are within proposed safe limits, and red areas are beyond these limits [1]

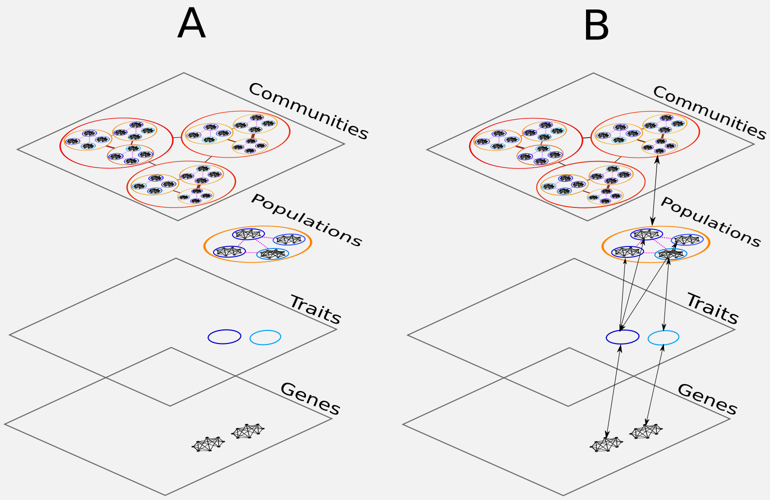


Figure 2: Biodiversity is sustained by processes acting across biological levels. Yet, inferring interdependencies among levels to predict the consequences of biodiversity decline remains poorly studied. A) Biodiversity has been studied mostly considering independent levels: Genes, traits and populations to communities and ecological networks each are defined at one level without exploring the interdependencies. B) Biodiversity represented as interdependent levels accounting for feedbacks among genes, traits, populations and communities. It remains unknown which of these two scenarios more accurately predict current trends in biodiversity decline and its consequences for Earth climate, life conditions and the stability of Earth.

2. Background and Significance to Data Science

The study of interactions, both within and across biological scales, is central to the ongoing synthesis of biodiversity [2, 3, 4]. Ecologists, for example, have argued for empirical patterns of positive [5], negative [6] and non-relationships [7] between the number of links and the stability of food webs (i.e., the number of links usually defined as the complexity of the food web). This debate is rooted in the mechanisms driving ecological interactions accounting for interactions among species [6, 8, 9, 10, 11, 12, 13].

Analogously, evolutionary biologists have puzzled over the relationships between the complexity of gene interactions and the stability of phenotypes tha drive ecological interactions [14, 15, 16, 17]. Quantitative genetics theory predicts that most genetic variance in populations is additive [18], and yet accounting for nonlinear gene interactions can improve predictions about the distribution and evolution of traits [19]. Experiments are also increasingly showing that gene interactions are common and that additivity can be an emergent property of underlying genetic interaction networks [20, 21, 22, 23, 24, 25]. Although the relationship between complexity and stability has been explored within biological levels, such as either genes, traits, populations or food webs, the interdependencies among levels have received less attention. Therefore, inferring interdependencies among biological levels to predict their role in biodiversity maintenance remains poorly studied [26, 27, 28, 29] (Figure 2).

Most methods in AI and Biodiversity research have been considered as distinct fields. However, the current scientific ecosystem is at the stage where merging methods from distinct fields is radically transforming the discipline boundaries, the reproducibility of science and our predicting-understanding power to build and test theories [30]. Recent approaches in ecology and evolution have introduced deep learning methods for labelled data, from which selection modes and demographic history can be jointly inferred [31]. Yet, many of the recent approaches applying deep learning methods in biodiversity have mostly focused at one level of biological organization. While this might produce additional gain in detailed knowledge at each level, it remains unknown how many layers are needed for predicting and understanding biodiversity patterns. The one-level and one-scale approach might be insuffiсent to understand the consequences of biodiversity decline.

To gain predictive and understanding power in biodiversity research we would need to develop deep process-based learning models accounting for many layers and the topology of the interactions within and between the layers [26, 27, 28, 29]. Many methods from data science and biological systems share fundamental properties, but the full potential of these shared properties have not been sufficiently explored [32]. Biological systems are composed by many layers (Figure 2), and they can contain interdependent hierarchies and feedbacks with interacting learning entities within and also between the layers. Therefore, exploring deep learning networks topologies accounting for feedbacks within and between layers is a first step towards understanding biodiversity dynamics using deep process-based learning networks (Box 2).

3. Scientific Goals and Objectives

Our goal is to study data-driven patterns within and between biological levels to decipher the role of interdependencies in maintaining biodiversity (Box 1 and 2). We will combine our own source data with other databases to develop deep process-based learning networks in biodiversity research. We will contrast two scenarios: An scenario without feedbacks within and between the biological levels and a second scenario accounting for feedbacks within and between layers to decipher the role of interdependencies across biological levels for biodiversity maintenance (Box 2). These two scenarios will be contrasted for a series of biodiversity patterns, mostly related to biodiversity maintenance and decline rates.

4. Research Plan

4.1 General description of the scientific approach

Our research plan is described in the following two boxes: Box 1 highlights the inference of our source empirical data to decipher patterns across biological levels. Box 2 describes deep process-based learning networks in biodiversity research.

4.2 Data sources

The following are our source datasets:

- **Projet Lac database**

  This is a dataset containing Fish communities from the majority of lakes of Switzerland (https://www.eawag.ch/en/department/fishec/projects/projet-lac/). The database contains approx. 60k fish individuals, 40k morphometrics and 8k sampling actions each containing spatial coordinates, morphological traits, species abundances, habitats and DNA data.

- **Whitefish community database**

  Whitefish community from the majority of lakes of North Alps. The database contains more than 2k fish individuals. Sampling locations, morphological, ecological and microsatellite data is available in the Dryad database (https://doi.org/10.5061/dryad.k183ft7).

- **Threespine Stickleback database**

  The Threespine stickleback fish database contains genetic, isotope, diet, morphometric, ecological and spatial data for around 1k individuals (Figure 3).

Box 1. Inferring patterns of interdependence in model systems

There are a few model systems with multidimensional data (section 4.2 Data Sources for a non-exhaustive list). We have performed a preliminary analysis of the Threespine Stickleback database containing genetic (QTL mapping), phenotypic (morphological), isotope, trophic (diets) and spatial (habitat) data for around 1k individuals. Our analysis show a normally distributed pairwise correlation pattern indicating a high degree of no correlations in the data (Figure 3a). A large cluster contains a high proportion of the individuals, but many small clusters and isolated individuals also occur (not shown) indicating high dimensionality of the data (Figure 3b: color indicates different clusters each containing similar individuals). We will explore an methods to infer the interdependence among genetic, phenotypic, isotope, trophic and the spatial dimensions of our source data [32, 33] (Figure 3c shows the target for this analysis with each level containing within- and between-layer interactions. Connections within and between layers are only for illustrative purposes.)

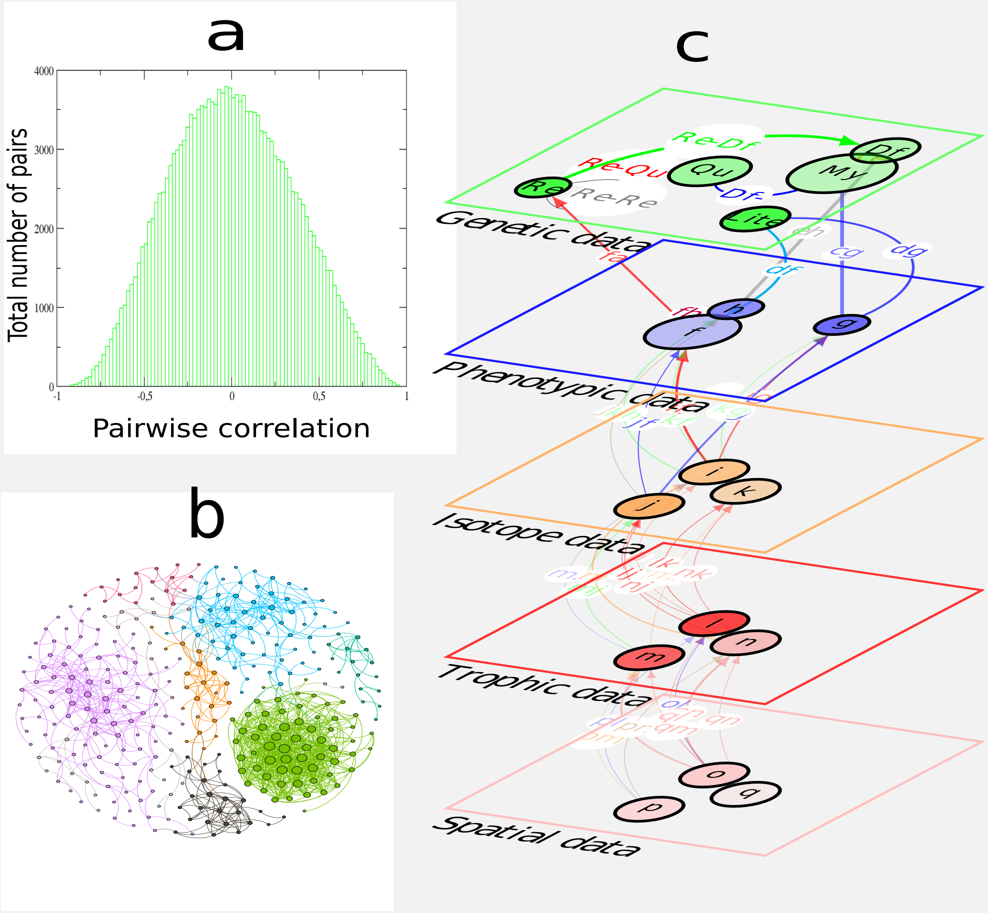
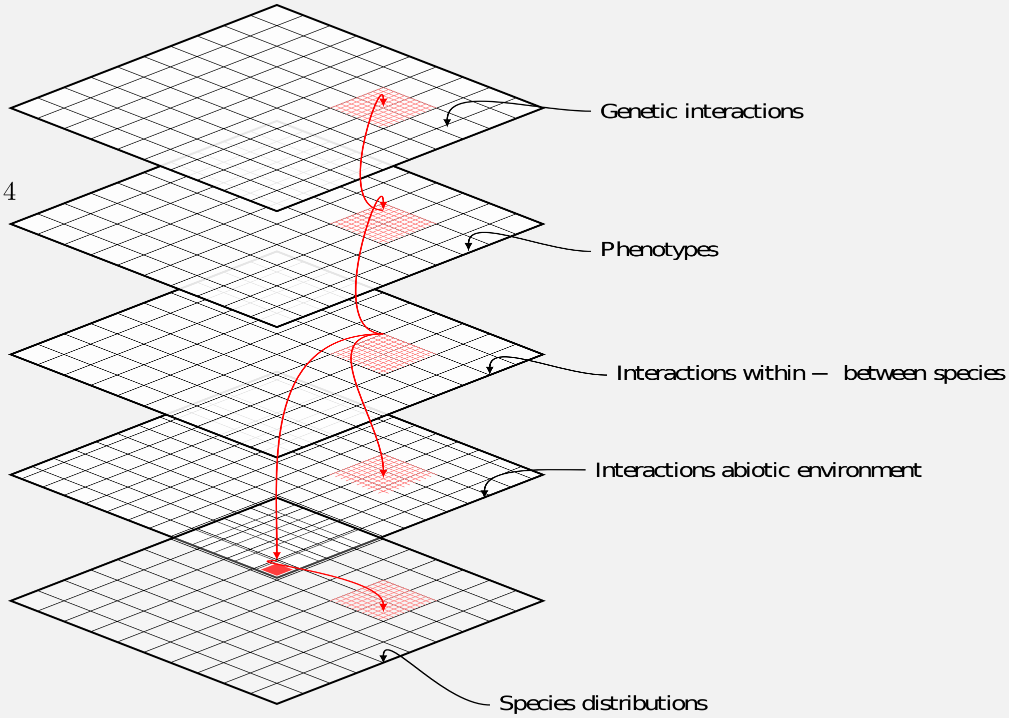


Figure 3

Box 2. Deep process-based learning networks in Biodiversity research

We will implement a multilayer approach to generate process-based species distribution maps accounting for interdependent biological networks (Figure 4). Each layer will be parametrized taking advantage of the empirical patterns obtained from our source data (Box 1) and from the integration of biodiversity datasets (See section 4.2 Data Sources and *https : //github.com/melian*009*/Robhoot/blob/master/layers/data.integration/databases.md.*) Most data in biodiversity are collections of small data. In areas such as species ranges and species interactions, there is a large amount of data, but only a relatively small amount of data for each gene, phenotype, individual or trophic interaction. To customize predictions accounting for interdependent biological levels it becomes necessary to build a formalism considering the heterogeneity at individual level, with its inherent uncertainties, and to couple these models together in a hierarchy scaling from genes, to phenotypes, populations, communities and species ranges, so that information can be borrowed from other similar levels across the landscape in the absence of empirical estimations. This individualization is becoming common in many fields [34]. We will implement a multilayer approach using hierarchical Bayesian neural networks such as hierarchical Dirichlet processes accounting for interdependent layers (Figure 4: Genetic-phenotype interactions determine the interactions within-between species and with the abiotic environment. The final result of the multilayer interactions will generate a biodiversity distribution map for many interacting species). We will explore a range of topologies from bidirectional recurrent neural networks (BRNN) to feedforward neural networks (FNN) and reinforcement learning in unknown and fluctuating environments (RL) [32]. We will explore independent networks considering modularity and sparse matrices within- and between-layers (i.e., a highly modular pleiotropy matrix determining the genotype-phenotype map and a highly modular within- and between-species interactions with most interactions weak or zero.) Such scenario will produce a non-interactive species biodiversity map. The second scenario will take into account interactions and feedbacks among layers to contrast predictions between independent and interdependent layers in biodiversity dynamics and how the interactions and feedbacks alter biodiversity decline to different disturbance regimes [29].

Figure 4


Genetic interactions
Phenotypes
Interactions within − between species
Interactions abiotic environment
Species distributions

4.3 Work Packages, Milestones and Deliverables

This project will strengthen the feedback between data-scientists and the community of scientists in biodiversity research. This fussion will produce two methods packages and two scientific papers. Below we describe the timeline describing the milestones, the deliverables and the timing to release the packages in public repositories.

Work packages
WP1: Data visualization
Lead: SDSC, contributed data: GENOM, ECOSYS and BIOD
WP2: Inference empirical patterns with our source data (Box 1 and Section 4.2)
Lead: SDSC, contribution: MAIN, BIOD, CSYS, GENOME and ECOSYS
WP3: Deep process-based learning networks (Box 2)
Lead: SDSC, MAIN and CSYS)

Milestones
M1.1: Uploading source data to the SDSC Platform
M1.2: Visualization empirical patterns of interdependent networks
M2.1: Web spidering to enrich the existing database in biiodiversity research
M2.2: Animation package of the Fish collection part of the FishEc database in the permanent exhibition hosted by the Natural History Museum in Bern
M3.1: Implementation of methods for inferring patterns in interdependent networks (Box 1)
M3.2: Efficient code implementation for inferring the interdependence among networks
M4.1: Implementing, debugging, running and analyzing the deep process-based learning networks scenarios (Box 2)
M4.2: Global biodiversity maps without and with interdependent networks

Deliverables
D1: Data visualization package to gain insights of the interactions across biological networks from genes to ecosystems.
The package will be used as highlights in a permanent exhibition of the natural history museum in Bern. We will produce a public github repository and a reproducible research document in Jupyter and Renku.
D2: Data mining and inference patterns and processes in interdependent networks.
We will produce the inference package in interdependent networks. The package will be uploaded and maintained in a github public repository. Together with the inference package, we will produce a reproducible research document in Jupyter and Renku.
D3: Scientific paper focusing on pattern inference in interdependent networks.
We will aim to a top general scientific journal. A reproducible research document will be developed in github, Jupyter and Renku.
D4: Scientific paper focusing on deep process-based learning networks introducing the global map of biodi-

versity in interdependent networks.
We will aim to a top general scientific journal. A reproducible research document will be developed in github, Jupyter and Renku.

5. Requested Resources

5.1 Staff

5.1.1 Data science expertise

A two years data scientist from the SDSC will be required to guarantee the accomplishment of the word packages WP1 to WP3, and the deliverables D1 to D4 (see Requested resources and contributed resources below). The following are the key contributions for each stage of the project:
WP1: Software development to complement/improve the existing visualization and animations tools for multilayer networks. Many libraries are rapidly emerging to integrate, analyze, and visualize patterns in multilayer networks, yet new features will be required to gain insights of the coupling of interacting biological networks.
WP2: Machine Learning techniques to enrich the biodiversity database.
WP3: Efficient code implementation to infer patterns of interactions in the empirical multilayer networks (Box 1).
WP4: Analysis of the deep process-based learning networks to generate biodiversity maps accounting for interdependent networks (Box 2).

5.2 Compute and storage resources

5.2.1 Summary

The following are the working packages requiring computing and storage resources:

WP1: Run the visualizations for a small dataset containing 5-10k individuals with genetic, phenotypic, ecological and spatial data. We will scale the visualization for this subset to a larger dataset containing approximately 60k fish individuals, 40k morphometric measurements, and 8k sampling actions each containing spatial coordinates, morphological traits, abundances, habitats and DNA data. We expect approximately 4 GPU during 6 months (see **SDSCfull.proposal.requested.resources.Melian.2019.xlsx**) to run the visualizations and the animations using the small and the large dataset and approx. 40 GPU servers RAM during a total period of 6-8 months.
WP2: Implementing Gibbs sampling to fit the interaction coefficients for the interactions within and between the genetic, phenotypic, ecological and spatial data (Box 1 and 2). Our estimated need for CPU (cores) is of 35 for a 8 months period. Please notice the fitting for Box 1 using our source data can be quick in comparison to the fitting to produce the biodiversity maps using deep process-based learning networks.

We expect to be working with the following data size for our source data: 1) Data containing 5-10k individuals each containing a 20x20 gene interaction matrix and a 40x40 phenotypic interaction matrix. The spatial data contains 0.1k-0.5k sites each containing between 1 and 10 habitats. Each individual has coordinates within a habitat.
2) Data containing 50k individuals each containing a 40x40 gene interaction matrix and a 80x80 phenotypic interaction matrix. The spatial data contains 0.5k-1k sites each containing between 1 and 10 habitats.
WP3: Implementation of the fitting in the deep process-based learning networks (Box 2). We will be working with matrices as described in WP2. Please notice the biodiversity maps will contain much larger matrices because the integration between different datasets. We expect the numbers described above will be two-three times larger.

5.2.2 Software packages

The following is the list of packages and skills required to accomplish the deliverables D1 to D4:
Skill 1: Spark GraphX to combine visualization, exploratory analysis and computation of metrics to infer network patterns crossing two or more networks.
Additional skills: Implementation of packages in python (pymnet), java (gephi), and julia (muxviz) languages.

Skill 2: Javascript/Jquery to enrich the biodiversity database.
Additional skills: Implementation of codes in python, Ruby or others to search databases.

Skill 3: TensorFlow. We have been working at a very preliminary stage with a julia wrapper for TensorFlow. We are open to learn from other packages/languanges to implement the methods required in this proposal.
Additional skills: Hidden random Markov fields models, Bayesian methods (Dirichlet process among others) and deep learning networks (bidirectional recurrent neural networks, feedforward neural networks and reinforcement learning networks) to efficiently compute and/or reconstruct interaction networks using genetic, phenotypic, ecological and spatial data.

6. Contributed resources

Dr. Victor Eguiluz, team CSYS, will join as a senior scientist during the spring 2020 to work in the work packages WP2 (inference) and WP3 (Process-based modeling). The main task of Dr. Eguiluz will focus on:
1. Implementing, debugging, running and analyzing the proposed deep process-based learning networks scenarios (Box 2).
2. Drafting scientific paper focusing on inference patterns in interdependent networks (Deliverables D3 and D4). Please see the document **SDSCfull.proposal.requested.resources.Melian.2019.xlsx** for a view of the total contributed staff.
Contributions for other groups:

MAIN: Dr. Carlos Melian: (1 x 4)
Teaming up with SDSC staff and Dr. Eguiluz (0.3 x 24) to develop pattern and process-based inference packages (Deliverables D2 to D4 with SDSC and work packages WP3 and WP4 with SDSC and Dr. Eguiluz).

2 months **GENOM** Dr. Philine Feulner: (0.2 x 2)
Supervising accuracy of the genetic data included in the FishEc dataset (WP1).
2 months **ECOSYS** Dr. Blake Matthews: (0.2 x 2)
Drafting scientific papers focusing on pattern inference and process-based methods in interdependent networks (deliverables D3 and D4)
2 months **BIOD** Prof. Ole Seehausen: (0.1 x 2)
Supervising accuracy of the genetic and morphological data in the FishEc dataset (WP1)
Supervision of the visualization package (D1) and organization of the permanent exhibition of the natural history museum in Bern
All the groups
Drafting scientific papers focusing on pattern inference and process-based methods in interdependent networks (deliverables D3 and D4).

References

[1] T. Newbold, L. N. Hudson, A. P. Arnell, S. Contu, A. De Palma, S. Ferrier, S. L. L. Hill, A. J. Hoskins, I. Lysenko, H. R. P. Phillips, V. J. Burton, C. W. T. Chng, S. Emerson, D. Gao, G. Pask-Hale, J. Hutton, M. Jung, K. Sanchez-Ortiz3, B. I. Simmons, S. Whitmee, H. Zhang, J. P. W. Scharlemann, and A. Purvis. Has land use pushed terrestrial biodiversity beyond the planetary boundary? a global assessment. *Science*, 353:288–291, 2016.

[2] C. Darwin. *On the Origin of Species.* Harvard Univ. Press, Cambridge, 1964.

[3] D. J. Futuyma and M. Slatkin. *Coevolution.* Sinauer Associates Inc. Sunderland, Mass., 1983.

[4] J. N. Thompson. *Relentless Evolution.* University of Chicago Press, Chicago, 2013.

[5] R. MacArthur. Fluctuations of animal populations and a measure of community stability. *Ecology*, 36: 533–536, 1955.

[6] R. M. May. *Stability and Complexity in Model Ecosystems.* Princeton Univ. Press, Princeton, USA., 1973.

[7] C. Jacquet, C. Moritz, L. Morissette, P. Legagneux, F. Massol, P. Archambault, and D. Gravel. No complexity-stability relationship in empirical ecosystems. *Nature Communications*, 7:12573, 2016.

[8] J. A. Dunne, U. Brose, R. J. Williams, and N. D. Martinez. Modeling food-web dynamics: complexity-stability implications. *Aquatic food webs: an ecosystem approach, Andrea Belgrano, Ursula M. Scharler, Jennifer Dunne, and Robert E. Ulanowicz (eds)*, pages 117–129, 2005.

[9] E. T. Thebault and C. Fontaine. Stability of ecological communities and the architecture of mutualistic and trophic networks. *Science*, 329:853–856, 2010.

[10] S. Allesina and S. Tang. Stability criteria for complex ecosystems. *Nature*, 483:205–208, 2012.

[11] Johnson S., Domínguez-García V., Donetti L., and Mu noz M. A. Trophic coherence determines food-web stability. *Proceedings of the National Academy of Sciences of the U. S. A.*, 111:17923–17928, 2014.

[12] A. Mougi and M. Kondoh. Food-web complexity, meta-community complexity and community stability. *Scientific Reports*, 6:24478, 2016.

[13] D. Gravel, F. Massol, and M. A. Leibold. Stability and complexity in model meta-ecosystems. *Nature Communications*, 7:12457, 2016.

[14] P. Alberch. From genes to phenotype: dynamical systems and evolvability. *Genetica*, 84:5–11, 1991.

[15] J. S. Arnold. Constraints on phenotypic evolution. *The American Naturalist*, 140:S85–S107, 1992.

[16] V. Debat and P. David. Mapping phenotypes: canalization, plasticity and developmental stability. *Trends in Ecology and Evolution*, 16:555–561, 2001.

[17] A. Wagner. *Robustness and evolvability in living systems.* (Princeton University Press), Princeton, USA. 2005.

[18] W. G. Hill, M. E. Goddard, and P. M. Visscher. Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genetics*, 4:e1000008, 2008.

[19] S. K. G. Forsberg, Sadhu M. J. Bloom, J. S., L. Kruglyak, and Ö Carlborg. Accounting for genetic interactions improves modeling of individual quantitative trait phenotypes in yeast. *Nature Genetics*, 49:497–503, 2017.

[20] F. W. Stearns. One hundred years of pleiotropy: A retrospective. *Genetics*, 186:767–773, 2010.

[21] A. Eyre-Walker. Genetic architecture of a complex trait and its implications for fitness and genome-wide association studies. *Proceedings of the National Academy of Sciences*, 107:1752–1756, 2010.

[22] G.P. Wagner and J. Zhang. The pleiotropic structure of the genotype-phenotype map: the evolvability of complex organisms. *Nature Reviews Genetics*, 12:204–213, 2011.

[23] T. F. C. Mackay. Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nature Reviews Genetics*, 15:22–33, 2014.

[24] L. North, T. and M. A. Beaumont. Complex trait architecture: the pleiotropic model revisited. *Scientific Reports*, 5:9351, 2015.

[25] M. Pavličev and J. Cheverud. Constraints evolve: Context dependency of gene effects allows evolution of pleiotropy. *Annual Review of Ecology Evolution and Systematics*, 46:413–434, 2015.

[26] T. G. Whitham, J. K. Bailey, J. A. Schweitzer, S. M. Shuster, R. K. Bangert, C. J. LeRoy, E. V. Lonsdorf, G. J. Allan, S. P. DiFazio, B. M. Potts, D. G. Fischer, C. A. Gehring, r. L. Lindroth, J. C. Marks, S. C. Hart, Wimp G. M., and S. C. Wooley. A framework for community and ecosystem genetics: from genes to ecosystems. *Nature reviews Genetics*, 7:510–523, 2006.

[27] N. Loeuille. Influence of evolution on the stability of ecological communities. *Ecology Letters*, 13: 1536–1545, 2010.

[28] C. Fontaine, P. R. Guimarães, S. Kéfi, N. Loeuille, J. Memmott, W. H. van Der Putten, F. J. F. van Veen, and E. Thébault. The ecological and evolutionary implications of merging different types of networks. *Ecology Letters*, 14:1170–1181, 2011.

[29] C. J. Melián, B. Matthews, C. S. Andreazzi, J. P. Rodrǵuez, L. J. Harmon, and M. A. Fortuna. Deciphering the interdependence between ecological and evolutionary networks. *Trends in Ecology and Evolution*, 33:504–512, 2018.

[30] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 566:195–204, 2019.

[31] S. Sheehan and Y. S. Song. Deep learning for population genetic inference. *PLoS Comput. Biol.*, 12: e10048452, 2016.

[32] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

[33] G Bianconi. *Multilayer Networks; Structure and Function.* Oxford University Press. Oxford. 2018.

[34] Z. Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521:452–459, 2015.